Research Journal of English Language and Literature (RJELAL)
A Peer Reviewed (Refereed) International Journalhttp://www.rjelal.com

Vol.3.3.2015 (July-Sep)

RESEARCH ARTICLE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2636 (Print):2321-3108 (online)

MEASURE FOR MEASURE: THE ESSENCE OF INCORPORATING RATING SCALE IN ASSESSING WRITING AT TERTIARY LEVEL

TAHMINA MARIYAM[1], Md. MOHIB ULLAH[2]

[1] Lecturer, Department of English Language and Literature, International Islamic University Chittagong, Chittagong, Bangladesh

[2]Assistant Professor, Department of English Language and Literature, International Islamic University Chittagong, Chittagong, Bangladesh

ABSTRACT

General impression marking, is said to be the less reliable predecessor of holistic scoring. If there is a predecessor to general impression marking, the Department of English Language and Literature of International Islamic University Chittagong seems to be using that! Out of our 35 core courses, 27 are related to literature; and thus require subjective scoring. Our testing system consists of a 30 marks' mid-term and a 50 marks' final-term. Sadly, we do not have any criteria, explicit or implicit, in order to judge the examination scripts. Addressing the issue, the paper tries to find out how intra-rater and inter-rater reliability is being hampered in the present testing system. It starts with the description of different types of rating scales and ends with a practical recommendation to lessen fluctuation in scoring.

Keywords: Writing Assessment, Rating scale, IIUC, Rater Reliability

1. INTRODUCTION

The present paper deals with the rating procedure of English writing at a Bangladeshi tertiary education institution (Department of English Language and Literature, International Islamic University Chittagong); "that is, in a second language adult education context" (Shamsuzzaman and Everatt, 2013: 69). Though the majority of the courses call for subjective assessment, no existence of a well developed rating scale is to be found in the department.

The Department of English Language and Literature has been doing without any rating guidelines since its inception in the year 2000. Whereas rating or scoring procedure is one of the most important aspects of any institution. As Weigle points out, "the scoring procedures are critical because the score is ultimately what will be used in making decisions and interferences about writers" (2002: 108).

McNamara asks us to 'imagine' two situations. First, "in which the ratings which candidates get depend not at all on the quality of their performances, but entirely on the whim of the rater". And the second, "the opposite case of the ideal rating system" he comments that the actual "situation will lie somewhere between these two extremes" (2000: 56-57). But in a situation where there is no guidelines regarding the rating procedure, the first imagination, that of a whimsical situation is bound to turn into reality. Sadly that

seems to be the case regarding the institution in focus in the paper.

The way out of invalid and unreliable scoring lies in that of a properly designed rating scale. As "the role of a scale is rather as a tool for raters to use, to help in channeling the diverse set of reactions raters have when they read texts into narrower, more manageable… statements about them" (Lumley, 2002: 268).

1.1. Statement of the Problem

Reliable and valid scores seem to be the expectation of anyone and everyone related to a certain test. A well developed rating scale does not only result in valid and reliable scores, rather it also calls for rater training resulting into more reassurance regarding assessment. As Lumley asserts, "rating is certainly possible without training, but in order to obtain reliable ratings, both training and reorientation are essential" (Lumley, 2002: 267).

The present scenario of the institution in question seems to be the exact opposite. Surviving without a rating scale for such a long time makes the validity and reliability of the assessment questionable. Also due to the lack of a scale the institution seems to lack a total synchronization between the raters, test takers and the institution.

A study that emphasizes the significance of rating scales could help finding out the situation and expectation of all the personnel related to the rating procedure of the present institution.

1.2. Purpose of the Study

The purpose of the present study is to show the necessity of a well developed rating scale in order to achieve inter-rater and intra-rater reliability. It also intends to show how the students are willing to attain a specific scoring guideline. It also tries to find out the attitude of the teachers towards rating procedures.

In doing so the study follows an approach that combines both qualitative and quantitative data analysis. After eliciting data from the students, teachers and assessed scripts the study shows the necessity of a common rating scale. The study ends with a recommended adaptable rating scale and some concluding remarks.

2. Review of the Literature:

Rating scale is a type of scale which consists of several ranked and structured categories used for making assessment. In rating scales of assessment, there are band descriptors which clarify the interpretation. Nunn states, "A rating scale is a practical means of assessing the level of a particular communicative performance by using a number of descriptive bands for a particular skill on a scale of competence ranging from excellence to failure" (2000: 171). McNamara outlines that a rating scale is used while assessing learners' performance by the assessors and it is "an ordered set of descriptions of typical performances in terms of their quality" (2000: 136). McNamara (1996) further states that the proper design, development and description of the scale are very significant for the validity of assessment. Upshur and Turner mention, "Although ratings have been regularly used in modern second language teaching, systematic concern with the development and characteristics of second language dates from 1970s" (1995: 4).

Rating scale is helpful to teacher, teacher-assessor and learners. Rubrics in rating scale help teachers to teach learners, focusing on issues that can carry good marks for the learners. Besides, rating scale is supportive for the assessors in assessing the writing tasks of students. More importantly, test takers can have clear idea from the rubrics of rating scale about the factors, maintaining which, can ensure good grades for them. Nunn argues,

> Rating scales also focuses attention on what both individual students and groups of students are good at, and what needs more attention while there is disagreement about the significance of 'washback effect' of test, it does not seem controversial to suppose that in graded courses in institutional learning, what is seen to be tested is more likely to be taught and learnt. Descriptions of desired performances link the students' natural ability to pass exams to the need to develop real language skills. (2000: 171).

Alderson (1991, cited in Nunn, 2000) discusses the reasons for using rating scales in some details, but only a short summary will be provided here. Firstly, rating scales provide an easily understandable

report (op. cit: 72) for candidates, administrators, course designers, and teachers on the level of performance of individuals or groups, at the same time as providing descriptions of what candidates can do. They can report on 'typical or likely behaviours of candidates at any given level' or on the proportions of candidates at each level. Secondly, rating scales can guide the rating process (op. cit: 73) standardizing the criteria for an individual rater or act as 'a common standard for different rates'. Finally, they also help to guide the construction of tasks (op. cit: 74) which allow students to display the described behaviours at their own level.

Underhill in this connection states rating scales are,

" … a series of short descriptions of different levels of language ability. The purpose of the scale is to describe briefly what the typical learner at each level can do, so that it is easier for the assessor to decide what level or score to give each learner in a test. The rating scale therefore offers the assessor a series of prepared descriptions, and she then picks the one which best fits each learner" (1987: 98)

According to Hughes (2003), there are two kinds of rating scales- Holistic/Global scale and Analytical Scale. He says, "Holistic scoring (sometimes referred to as impressionistic scoring) involves the assignment of a single score to a piece of writing on the basis of an overall impression of it" (p.94). While defining Analytical scale, he states, "Methods of scoring which require a separate score for each of a number of aspects of a task are said to be analytic" (p.100).

Davies et al. (1999, cited by Nakamura, 2002)) opine that in the assessment of writing, a major advantage of holistic over analytical scoring is that each writing sample can be evaluated quickly by more than one rater for the same cost that would be required for just one rater to do the scoring using several analytic criteria.

In 1996, McNamara presented different factors that consist of a typical performance assessment.

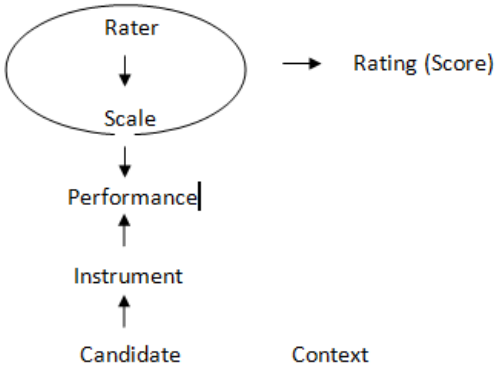

Figure 1. Factors in performance assessment (Adapted from McNamara, 1996)

Arguable though, it is observed that rating scales are not always efficient and effective, for there might be big issues of reliability and validity in commonly used rating scales. (Bachman and Savaignon 1986, Fulcher 1987, Matthews 1990). Context and culture more or less influence the writing of the learners and so they should also be taken into consideration in assessing writing. In 2002, Weigle argued, "the implication for the testing of writing is that writing ability cannot be validly abstracted from the contexts in which writing takes place." Weir and Shaw (2008) outline context validity as a crucial element in maintaining stand and worth of writing tests. The following figure shows the aspects of scoring validity in writing:

Scoring Validity
Criteria/rating scale
Rater Characteristics
Rating process
Rating conditions
Rater training
Post-exam adjustment
Grading and awarding

Figure 2. Criteria for scoring validity for writing (adapted from Weir 2005, p. 47)

Ghanbari, Barati and Moinzadeh's work on the rating scales in EFL academic writing assessment in the context of Iran. Their findings "indicated a vacancy for an objective measure in EFL writing assessment" (2012: 86). They recommended for training of the raters and developing a local rating instrument in Iranian context to solve the problem.

Also, we notice some research has been conducted on assessing writing and rating scale in the context

of Bangladesh. Kabir (2007) works on the validity and reliability in testing two skills- reading and writing at the Higher Secondary level. His findings show that there are lacking in validity and reliability in testing the mentioned skills. He suggests for training of the raters, question setters and improvement of test administration situation to make assessment more reliable and valid. It is important to initiate "a feasible rating scale" to assess writing skill more reliably at the HSC level (Kabir, 2012).

3. Research Design

Keeping the purpose of the study in mind a descriptive research design has been selected for the present research. As regarding descriptive research Seliger and Sohamy assert, "It is similar to qualitative research… in addition, descriptive research is often quantitative" (2003: 124).

3.1 Participants and Setting:

Teachers and students of the Department of English Language and Literature of International Islamic University Chittagong are the participants in the present study.

3.2. Data Instruments:

Three data instruments have been used in order to conduct the study. It includes: questionnaire for students, questionnaire for teachers and assessment of scripts.

3.2.1. Questionnaire for students

A questionnaire, with seven statements and four probable responses, was distributed among 25 students in order to find out their attitude toward the existing rating system and the importance of incorporating a rating scale for better assessment.

3.2.2. Questionnaire for teachers

A questionnaire in order to elicit both qualitative and quantitative data regarding their attitude towards rating and rating scale was distributed among five teachers.

3.2.3. Assessment of Scripts

11 midterm scripts (full marks: 30) were assessed by three different raters in order to find out inter-rater reliability. And again 5 midterm scripts (full marks: 30) were assessed by the same rater in two different occasions in order to find out intra-rater reliability.

4. Data Analysis

4.1. Response to the questionnaire for students

25 questionnaires were distributed among 25 students of different semesters. All of them were returned. The questionnaire consisted of seven statements with four grids of possible response. The response has been summed up below:

Table 1: A Summary of the Students' Response to the Questionnaire

Statements	Response 1	Response 2	Response 3	Response 4
	Always	Sometimes	Never	Not Sure
1. Ability of interpreting the attained scores	16%	72%	12%	0%
2. Willingness to know the reason behind the marks received	40%	48%	8%	4%
3. Questioning the teacher regarding the marks achieved	4%	64%	28%	4%
4. Getting satisfactory answers from the teacher regarding marks	36%	52%	4%	8%
5. Willingness to know the criteria of good writing that will bring good marks	80%	20%	0%	0%
6. Willingness to know the reason behind getting less marks than a friend despite of writing similar answers	8%	64%	20%	8%
7. Feeling the need of proper scoring guidelines	92%	4%	0%	4%

4.2. Response to the questionnaire for teachers

Five questionnaires were distributed among randomly selected faculty members of the department. All of them were filled and duly returned. The questionnaire consisted of three parts.

The first section was regarding personal information. It revealed that the respondents' teaching experiences ranged from four years to twenty years.

The second part consisted of fifteen statements along with five grids of possible responses. The response has been summed up below:

Table 2: A Summary of the Teachers' Responses to the Questionnaire

Statements Concerning	Responses
1. Following a personal way of scoring and claiming that their impressionistic scoring method is quite trustable. (Questions 1-5)	Almost all the respondents either agreed or strongly agreed. Only a single respondent disagreed.
2. Acknowledging the significance of rating scales regarding ensuring valid and reliable scores. (Questions 6-7)	Almost all the respondents either agreed or strongly agreed. Only a single respondent claimed to be uncertain.
3. Knowledge regarding different rating scales, inter-rater and intra-rater reliability. (Questions 8, 11, 15)	The responses turned out to be mixed. Some showing a lack of knowledge while others seemed to have proper knowledge. And some others claimed to be uncertain.
4. Acknowledging how the lack of a common rating scale leads to biased and inconsistent scoring; whereas the existence of one would result into a fair assessment.	All the respondents either agreed or strongly agreed to the statements.
5. Feeling the need of either a locally developed or an adopted international rating scale. (Questions 14, 10)	All the respondents felt the need.

The third part consisted of three open ended questions.

Question No.16 asked about important dimensions in students' writing. The responses were varied. Some considered content, vocabulary and organization to be the most important ones, because "without those the writing would appear to be devoid of any sense". Others considered cohesion and syntax to be of utmost significance as "those are the main modes of expressing ideas".

Question No. 17 asked about the purpose behind assessing students' writings. All the respondents unanimously agreed the purpose to be "identifying students' weakness and strength". Such diagnostic information "acting as direct feedback for the students" will pave a way towards "their gradual improvement".

Question No. 18 asked if the incorporation of a rating scale will influence the way they score. All of them answered on the affirmative for varied reasons.

According to T1, "it will bring consistency and reliability among all the scorers". T5 also envisioned similar output, as "the rating scale will guide me how to assess reliably".

For T2, "it will result into unified judgment regarding scores".

T3 said "it would increase the effectiveness of teaching and learning".

T4 commented that, "it will reduce the tension and anxiety in ensuring consistency regarding assessing the scripts of the students".

4.3. Assessment of Scripts

The scores obtained from three different raters after their assessment of 11 scripts have been presented below:

Table 3: Scores Obtained from Inter-rater Assessment

Rater 1	Rater 2	Rater 3
16.75	17	14
5	3	2
14	13.5	14.5
15	18	18
15	18	13
14	15	15.5
9	8	12
19.5	20	20
13	12	15

The scores obtained from a single rater after her assessment of 5 scripts on two different occasions have been presented below:

Table 4: Scores Obtained from Intra-rater Assessment

Occasion 1	Occasion 2
13	11
11	14
3.75	3
15	14
15	17

5. Findings and Discussion

The questionnaire for students showed their inability to interpret the scores they receive, on a regular basis. It also portrayed their eagerness to know the reasons behind the scores they receive. Very often they do present their queries to the course teacher regarding their obtained scores. It is not always that they receive satisfactory answers from the course teachers. They showed a willingness to know what kind of writing will result into higher scores. Very often they encounter a mismatch between the scores of the candidates. Almost all of them asked for explicit guidelines regarding scoring.

The questionnaire for teachers exposed a number of things. It revealed how all of the respondents follow their own way of scoring due to the lack of one. It shows some of the respondents' lack of knowledge regarding different types of rating scales as well as inter-rater and intra-rater reliability issues. Almost all of them agreed that biased and inconsistent scores are the direct result of the lack of a common rating scale. Acknowledging the significance of rating scale all of them showed the willingness of incorporating a rating scale for assessing writing, either locally developed or adopted from an international one.

The assessment of scripts showed how both inter-rater and intra-rater reliability is being hampered due to the absence of a rating scale. In very few occasions students received the almost the same score from three different raters. Even the scores varied a good deal when assessed by the same rater at two different occasions. As in the IIUC grading system is such where grades vary within five marks' difference, this scenario results in flawed judgment regarding assessing writing. Thus it won't be surprising if a strong student scores less than that of a weak one due to inter-rater or intra-rater inconsistency.

6. Recommendation

After observing the entire scenario it seems clear that a well developed rating scale is the solution to the problems related to scoring writing. Keeping the findings in mind the researchers recommend a rating scale adopted from TOFEL writing scoring guide. The recommended scale addresses the essential aspects of a writing task: content, organization, mechanics, syntax etc. The scale has been presented below:

Table 5: Recommended Rating Scale (Adopted from TOFEL writing scoring guide).

Marks to be assigned out of 10	Descriptors
8-10	Effectively addresses the writing task. Is well organized and well developed. Uses appropriate details to support ideas. Demonstrates syntactic variety and appropriate word choice (may have occasional errors).
7-6	May address some parts of the task more effectively than others. Generally well organized and developed. Uses details to support an idea. Demonstrates some syntactic variety and range of vocabulary (with occasional errors)

5-4	Inadequate organization or development. Inappropriate or insufficient details. Noticeably inappropriate choice of words. An accumulation of errors in sentence structure and/or usage.
3-2	Serious disorganization or underdeveloped. Little or no detail. Irrelevant specifics. Serious and frequent errors. Serious problems with focus
1-0	Incoherent. Undeveloped. Severe and persistent writing errors. Merely copies the topic. Is off-topic.

7. Conclusion

The study seems to have portrayed the reality of a testing context that tries to do without a proper rating scale. Pointing out the need of one, the researchers have recommended a rating scale which can be adapted to suit the local needs.

Though the limitation of the study lies in the fact that it deals with one institution, again this very fact seems to possess the ability of being turned into its strength. More single studies like that of the present one will lead into a database of the assessment done in different tertiary level institutions in Bangladesh. Thus from the findings of the study proper measures may be taken in order to measure or assess the tertiary level writing in English successfully.

References

Alderson, J C. "Bands and Scores." *Language Testing in the 1990s: The Communicative Legacy*. Ed. J C Alderson and Brian North. London: Modern English Publications in association with the British Council, 1991. 71-85. Print.

Bachman, Lyle F., and Sandra J. Savignon. "The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview." *Modern Language Journal*10.4 (1986): 380-90. Print.

Fulcher, Glenn. "Tests of oral performance: the need for data-based criteria." *ELT Journal* 4.14 (1987): 287-91. Print.

Ganbari, Batoul, Hossein Barati, and Ahmad Moinzadeh. "Problematizing rating scales in EFL academic writing assessment: voices form Iranian Context." *English Language Teaching* 5.8 (2012): 76-90. Web. 17 Mar. 2015. <http://www.ccsenet.org/journal/index.php/elt/article/download/18617/12334>.

Hughes, Arthur. *Testing for Language Teachers*. Cambridge [England: Cambridge UP, 2003. Print.

Kabir, Mohammed H. "An investigation into the validity and reliability in testing reading and writing skills at HSC level." *IIUC Studies* 3 (2007): n. pag. Print.

Kabir, Mohammed H. "Necessity of initiating rating scale for more reliable assessment of writing skill at HSC level: a case study." *IIUC Studies* 6 (2012): 35-51. Print.

Lumley, Tom. "Assessment Criteria in a Large-scale Writing Test: What Do They Really Mean to the Raters?" *Language Testing* 19.3 (2002): 246-276. Web. 25 Dec. 2014. <http://ltj.sagepub.com/content/19/3/246.abstract>.

Matthews, Margaret. "The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations." *ELT Journal* 44.2 (1990): 117-121. Print.

McNamara, T F. *Language Testing*. Oxford [England: Oxford UP, 2000. Print.

McNamara, Tim. *Measuring Second Language Performance: A New Era in Language Testing*. New York: Longman, 1996. Print.

Nakamura, Yuji. "A comparison of holistic and analytic scoring methods in the assessment of writing." *The interface between inter language. Pragmatics and assessment*. Proceedings of the 3rd Annual JALT Pan-SIG Conference. N.p., 2002. Web. 22 Nov. 2014. <https://jalt.org/pansig/2004/HTML/Nakamura.htm>.

Nunn, Roger. "Designing rating scales for small-group interaction." *ELT Journal*54.2 (2000): 169-178. Print.

Seliger, Herbert W., and Elana Shohamy. *Second Language Research Methods*. New York: Oxford UP, 2003. Print.

Shamsuzzaman, Mohammad, and John Everatt. "Teaching Writing in English at Tertiary Level in Bangladesh: Deconstructing error and reconstructing pedagogy."*Research and Educational Change in Bangladesh*. Ed. Janinka Greenwood, John Everatt, Ariful H. Kabir, and Safayet Alam. Dhaka: Dhaka Viswavidyalay Prakashana Sangstha, 2013. 69-84. Print.

Shaw, Stuart, and Cyril Weir. "Examining Writing: Research and Practice in Assessing Second Language Writing." *Studies in Language Testing 26* 11.4 (2008): xiv + 344. Web. 4 Feb. 2015. <http://www.tesl-ej.org/wordpress/issues/volume11/ej44/ej44r3/>.

Underhill, Nic. *Testing Spoken Language: A Handbook of Oral Testing Techniques*. Cambridge [Cambridgeshire: Cambridge UP, 1987. Print.

Upshur, J. A., and C. E. Turner. "Constructing Rating Scales for Second Language Tests." *Elt Journal* 49.1 (1995): 3-12. Print.

Weigle, Sara C. *Assessing Writing*. Cambridge: Cambridge UP, 2002. Print.

Weir, Cyril J. *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan, 2005. Print.

Appendix I

Questionnaire for Students

Statements	Option 1	Option 2	Option 3	Option 4
	Always	Sometimes	Never	Not sure
1. Are you able to interpret the scores you receive in your exams?				
2. Do you want to know the reason behind the scores/marks you receive?				
3. Do you question your course instructor regarding the marks/scores you receive?				
4. Do you get satisfactory answer from the instructor?				
5. Do you want to know, what kind of writing will bring more marks?				
6. Did you ever feel that you and your friend wrote the same thing, but she got more marks than you did?				

7. Do you need any scoring guideline?				

Appendix II

Questionnaire for Teachers

Dear Respondent,
Your cooperation for answering the following questions would be highly appreciable.

Gender: ☐ Male ☐ Female
Age: ……
Academic Qualification: ……
Area of specialization: ……
Teaching experience: …… Years

Statements	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
1. I follow my own way of scoring, thus never felt the need of any rating scale.					
2. I go through the text and based on my own experience in rating, I give a total score.					
3. Giving a score based on my impression is quite trustable.					
4. All raters have some criteria for their scoring though they might differ from each other.					
5. Upon experience, I have learned to keep all the rating criteria in my mind and score based on them.					
6. Rating scale plays a significant role in assessing writing.					
7. An explicit rating scale would improve validity and reliability of my assessment.					
8. Rating an answer is quite an individual act, there is no need of inter-rater agreement.					
9. Students are informed about my rating criteria early in the course.					
10. A local rating scale for writing assessment is needed to assure the validity and reliability of the scores.					

11. As a rater, I am quite aware of different rating scales.					
12. Lack of a common rating scale would lead to bias, inconsistency and leniency/severity among the raters.					
13. The existence of a common rating scale would lead to a more fair writing assessment					
14. We should adopt an international rating scale for fair assessment.					
15. My experience enables me to score the same at different occasions.					

16. Which dimension in students' writing is more important to you? Why? (You can choose more than one).

a. content ☐ b. vocabulary ☐
c. organization ☐ d. mechanics (spelling, punctuation) ☐
e. cohesion ☐ f. syntax ☐

Because……
…….

11. What do you think should be the purpose of assessing students' writings? Why?

a. Giving score ☐ b. Identifying students' weakness and strength ☐

Because……
……

12. Do you think incorporation of a rating scale will influence the way you score? Why?

Yes, Because ……

……

No, because
……
……

Appendix III

IIUC Grading System

Percentage of Score	Letter Grade	Quality Point
80-100	A+	4.00
75-79	A	3.75
70-74	A-	3.50
65-69	B+	3.25
60-64	B	3.00
55-59	B-	2.75
50-54	C+	2.50
45-49	C	2.25
40-44	D	2.00
00-39	F	0.00

Appendix IV

IIUC Marks Distribution for Theoretical Course (100 Marks)

Class Tests/Assignments (at least two) = 10 Marks

Class Attendance = 10 Marks

Mid-term Examination = 30 Marks

Final Examination = 50 Marks

(Recommendations from the discussions regarding academic curriculum between Honorable Pro Vice-Chancellor and Dean/ Heads/ Director of CENURC held on 14.05.12 and 05.06.12, IIUC: p2)